On the Application of Saddle-Point Methods for Combined Equilibrium Transportation Models

Demyan Yarmoshik[1,2][0000−0003−1912−1040] and Michael Persiianov[1,2,3][0009−0008−7059−1244]

[1] Moscow Institute of Physics and Technology, Dolgoprudny, Russia
{yarmoshik.dv,persiianov.mi}@phystech.edu
[2] Institute for Information Transmission Problems, Moscow, Russia
[3] The Skolkovo Institute of Science and Technology, Moscow, Russia

Abstract. Travel demand modeling is an essential tool in urban planning and transportation system management. Existing practically efficient algorithms for solving multistage travel demand problems are variations of the heuristic sequential procedure. We propose a novel approach that applies saddle-point methods to a combined convex optimization formulation of the problem. Unlike all previous methods, our algorithm does not require costly shortest-paths calculations, and can be seamlessly scaled on GPUs. We show that in some cases it drastically outperforms the sequential procedures.

Keywords: Convex optimization · Trip distribution · Traffic assignment · Combined transportation model · First-order methods · Affine constraints.

1 Introduction

In today's rapidly evolving urban landscapes, understanding and forecasting travel demand play pivotal roles in shaping efficient transportation systems and sustainable urban planning. Essentially, the purpose of a travel demand model is to predict the equilibrium traffic matrix and distribution of passenger and cargo flows through the transportation network, based on the network's characteristics. This information enables authorities to forecast the impact of changes in the transportation network, thus facilitating more informed decision-making.

The conventional four-stage approach [17], [13]: trip generation (travel choice), trip distribution (destination choice), modal split (travel mode choice), and traffic assignment (route choice) stages, a widely used method in travel-demand modeling, has long been the cornerstone of transportation planning and policy-making. However this approach, despite its prevalence, has a number of limitations, e.g. there is no convergence guarantee [12], [51], [7] due to tendency to overlook consistent consideration of travel times and congestion effects across its various stages [24], [45]. An alternative strategy to tackle this challenge involves adopting a mathematical model formulated as an optimization problem that can subsequently be solved via numerical optimization methods. Such models are commonly referred to as *combined* or *integrated* models.

1.1 Combined Models

For decades, there has been extensive research into models that combine various stages. The pioneering work [6] introduced the first constrained convex optimization formulation for the user-equilibrium assignment problem with elastic demand. This formulation treated travelers between every origin-destination (OD) pair as a function of the travel service for that specific OD pair. Building upon this foundation, [21] and [19] expanded the convex optimization formulation to include destination choice through a combined distribution and assignment (CDA) model. In this model, trip distribution follows a gravity model with a negative exponential deterrence function, while traffic assignment adheres to a user-equilibrium model. To account for congestion effects at destinations, [50] introduced variable destination costs into the CDA model.

Subsequent research extended the CDA model in various directions. [37], [9], and [29] refined the CDA model to accommodate multiple user classes, while [20] incorporated modal split within the CDA framework. [23] proposed an equivalent optimization problem for combined multiclass distribution, assignment, and modal split, eliminating symmetry restrictions. [60] put forth an optimization model that combines distribution, hierarchical mode choice, and assignment networks with multiple user and mode classes. [2] developed a comprehensive combined model by integrating all four steps of sequential demand forecasting based on random utility theory. [51] further extended this by simultaneously considering travel-destination-mode-route choice, utilizing the multinomial logit model within a hierarchical structure where each traveler is perceived as a customer of urban trips, with choices influenced by utility and budget constraints. Recent study [34], employing the entropy/utility maximization framework, proposed a convex programming approach equivalent to the hierarchical extended logit model which leads for the simultaneous estimation of all travel levels, thereby overcoming the input-output discrepancies in each level unlike traditional models with "feedback" loops [53].

1.2 CDA Model Applications

Among the discussed combined models, the CDA model holds particular significance in numerous transportation applications. It aids transportation planners in understanding the intricate interactions between land use and transportation, informing future urban development and transportation system improvement plans. For instance, researchers have utilized the CDA model within a bilevel programming framework to address various land use and transportation concerns. [62] applied the CDA model to analyze the capacity and level of service of urban transportation networks, while [56] used it to determine the maximum number of cars given road network capacity and available parking spaces. This model has been utilized to estimate the capacity of urban transportation networks incorporating rapid transit [16] and to identify critical links within road networks [18]. Besides, [15] introduced network-based accessibility measures to assess the vulnerability of degradable transportation networks, leveraging the

CDA model. Additionally, [42] employed the CDA model in a land use-network design problem to analyze integrated layouts of land uses, public facilities, transport networks, and travel demands. [40] investigated equity issues associated with land use development in terms of changes in equilibrium OD travel costs, while [63] used the CDA model to optimize network reliability concerning residential and employment allocations and network enhancements. Furthermore, [28] embedded the CDA model of housing locations and traffic equilibrium to determine optimal housing provision patterns in a continuum transportation system.

1.3 Solution Algorithms

When Beckmann et al. [6] initially proposed in 1956 the formulation of a model of elastic demand and route choice, no solution algorithm was provided. Twenty years later, Ph.D. student Suzanne Evans [19] introduced a partial linearization method for CDA problem so that only the objective function was linearized, and the OD choice functions was unchanged. While [21] and [20] presented another algorithm based on the Frank and Wolfe [22] method. However, according to [11] and [39], Evans' algorithm exhibits superior convergence characteristics.

Founded on two unproven conjectures, [30] introduced a modification to the Evans' algorithm aimed at reducing computational times and memory requirements. However, as highlighted by [31], Horowitz's adapted algorithm doesn't consistently converge to the optimal solution. In response, [31] proposed further refinements to Horowitz's approach, providing rigorous proof of convergence.

On the other hand, [44] proposed a hybrid solution method, merging the Evans' algorithm with the disaggregate simplicial decomposition (DSD) method developed by [38]. Expected to offer enhanced computational efficiency and greater solution accuracy, this method operates within the path-flow domain. However, combining the Evans' algorithm with the second-order DSD algorithm may lead to non-convergence.

More recently, [4], [5] extended the origin-based (OB) algorithm [3] for traffic assignment for tackling the CDA model as a fixed-point problem. This algorithm replaced the link-based assignment Evans' algorithm with an origin-based procedure to update the solution of the assignment problem, which demonstrated superior effectiveness in achieving highly accurate solutions. Notably, during the implementation of the OB algorithm, it was observed that the determination of step-size in the OD flow update step significantly impacts the algorithm's performance. To enhance the efficiency of the OB algorithm, [61] adopted the modified line search strategy proposed by [37] for the Evans' algorithm. In contrast, the work of [35] instead of using OB algorithms for convergence speed up utilized more prominent primal-dual methods [48] for traffic assignment sub-problem and [26] for trip distribution.

All previously mentioned algorithms can be divided into two groups: link-based (e.g., [22], [19], [35]) and origin-based (e.g., [4], [5]). *Link-based* algorithms prioritize aggregate link flows over individual path flows, making them memory-efficient and suitable for early computing environments. While they are easier to parallelize and implement, they can be slow, especially when high precision is required, due to convergence challenges. *Origin-based* algorithms focus on path flow vectors, offering faster convergence by retaining more information. Despite their higher memory demand and coding complexity, they excel in precision-critical scenarios by exploiting the vast number of paths in a network.

In this paper, we apply a convex optimization algorithm from [54] to solve the CDA problem. This approach offers several advantages over existing methods:

- The algorithm does not require solving the costly shortest-paths subproblem
- A single algorithm is used to solve the entire CDA problem, simplifying implementation and reducing the hustle of parameter tweaking
- The computation-heavy operations are element-wise function calculations and matrix-vector multiplication, which allows to write high-performance code in high-level programming languages like Python and seamlessly employ GPU acceleration via modern powerful frameworks such as [14].

The rest of the paper is organized as follows: Section 2 provides notation and mathematical model background. The algorithm for the doubly constrained CDA problem is presented in Section 3. Experiments are demonstrated in Section 4. Concluding remarks are discussed in Section 5.

2 Preliminaries

2.1 Notation

By $\sigma_{\max}(A), \sigma^+_{\min}(A)$ we denote maximal and minimal positive singular values of a linear operator A respectively. $\lambda_{\max}(A), \lambda_{\min}(A)$ denote maximal and minimal eigenvalues of a linear transformation A respectively. We use $\langle x, y\rangle := \sum_{i=1}^d x_i y_i$ to define the inner product of $x, y \in \mathbb{R}^d$. $\|x\| := \sqrt{\langle x, x\rangle}$ is the Euclidean norm of $x \in \mathbb{R}^d$, and $\|X\| := \sqrt{\langle X, X\rangle} = \sqrt{\mathrm{Tr}\,(X^\top X)}$ is the Frobenius norm of $X \in \mathbb{R}^{m\times n}$, where $\mathrm{Tr}(X)$ denotes the trace of X. By $\mathrm{Proj}_X(x_0) := \arg\min_{x\in X} \|x - x_0\|$ we mean Euclidean projection of x_0 on a set X. $1_{m\times n} \in \mathbb{R}^{m\times n}$, $1_n \in \mathbb{R}^n$, $0_{m\times n} \in \mathbb{R}^{m\times n}$, $0_n \in \mathbb{R}^n$ stand for matrices and vectors of all ones and zeros respectively. Indenity matrix is denoted by I_n. For $X \in \mathbb{R}^{n\times n}$ we define $\mathrm{diag}(X) \in \mathbb{R}^n$ to be the vector of diagonal elements of a matrix X; and if $x \in \mathbb{R}^n$, then $\mathrm{diag}(x) \in \mathbb{R}^{n\times n}$ is defined to be the diagonal matrix, constructed from a vector x.

2.2 Traffic Assignment. Beckmann Model

The traffic assignment problem is to find the equilibrium flow and travel time for each road given traffic matrix and traffic network. Following the outline of [35], let the traffic network be represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $|\mathcal{V}| = n$, $|\mathcal{E}| = m$, where vertices $\mathcal{V}$ correspond to intersections or centroids [55] and edges $\mathcal{E}$ correspond to roads. Suppose we are given the traffic matrix d of travel demands: namely, let d_{ij}(veh/h) be a trip rate from source $i \in \mathcal{V}$ to destination $j \in \mathcal{V}$. Here, to simplify the notation, we assume without loss of generality that all the vertices are sources and targets, and some of them can have zero travel demand or supply. We denote by P_{ij} the set of all simple paths from source $i \in \mathcal{V}$ to target $j \in \mathcal{V}$, and $P = \bigcup_{i,j\in\mathcal{V}\times\mathcal{V}} P_{ij}$ the set of all paths between sources and targets.

Agents traveling from node i to node j are distributed among paths from P_{ij}, i.e. for any $p \in P_{ij}$ there is a flow $x_p \in \mathbb{R}_+$ along the path p, and $\sum_{p\in P_{ij}} x_p = d_{ij}$. Flows from origin nodes to destination nodes create the traffic in the entire network $\mathcal{G}$, which can be represented by an element of

$$X = X(d) = \Big\{ x \in \mathbb{R}_+^{|P|} : \sum_{p\in P_{ij}} x_p = d_{ij};\ i,j \in \mathcal{V} \Big\}. \tag{1}$$

Then, the flow on edge e

$$f_e(x) = \sum_{p\in P} \delta_{ep} x_p \text{ for } e \in \mathcal{E},$$

where δ_{ep} equals 1 if $e \in p$ and 0 otherwise.

The problem of finding equilibrium load of a transportation network, known as traffic assignment problem, traditionally formulates [6] as

$$\Psi(x) = \sum_{e\in\mathcal{E}} \underbrace{\int_0^{f_e} \tau_e(z)dz}_{\sigma_e(f_e)} = \sum_{e\in\mathcal{E}} \sigma_e \left(\sum_{p\in P} \delta_{ep} x_p \right) \longrightarrow \min_{\substack{\sum\limits_{p\in P_{ij}} x_p = d_{ij}, \\ x_p \geq 0}}, \tag{2}$$

where $\tau_e(f_e)$ is a cost function, usually it is BPR function [58]

$$\tau_e(f_e) = \bar{t}_e \left(1 + \rho (f_e/\bar{f}_e)^{1/\mu} \right), \tag{3}$$

where $\bar{t}_e$ are travel times at zero flow, and $\bar{f}_e$ are road capacities of a given network's link e.

Since the number of all possible paths between two nodes in a typical transportation network is enormous, direct manipulations with vector of path flows x is unapproachable. Therefore, numeric solvers for traffic assignment problem typically use edge flows f_e as variable or apply column generation techniques [41]. These approaches rely on calculating shortest paths to make progress in minimizing Ψ while maintaining feasible edge flows w.r.t. traffic matrix d.

However, it is also possible to write down the constraint imposed by traffic matrix via Kirchhoff law. This formulation, which is commonly used for multicommodity flow problems [1], is our key technique to apply new saddle-point methods for CDA. Let us fix source $i \in \mathcal{V}$ and consider some node $j \in \mathcal{V}$. Define f to be the matrix of edge flows f from given source to all targets: the element f_{ei} of matrix f is the amount of flow originating at source i and traversing (oriented) edge e. Then, consider the flow originating from i that settles down in j. It writes as

$$\Delta_{ij} = \sum_{e\in\mathcal{E}_{in}(j)} f_{ei} - \sum_{e\in\mathcal{E}_{out}(j)} f_{ei}. \tag{4}$$

If $i \neq j$, this amount equals the flow running from i to j, i.e. $\Delta_{ij} = d_{ij}$. If $i = j$, this amount is the sum of flows originating from i with inverse sign, i.e. $\Delta_{ii} = -\sum_{k \neq i} d_{ik}$.

Further, let N denotes the *incidence matrix* of $\mathcal{G}$, i.e.

$$N_{ke} = \begin{cases} -1, & k = i, \\ 1, & k = j, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

for each edge $e \in \mathcal{E}$ with source $i \in \mathcal{V}$ and destination $j \in \mathcal{V}$. Note, that $(Nf)_{ji} = \sum_e N_{je} f_{ei} = \Delta_{ij}$. On the other hand, $\Delta_{ij} = -(\operatorname{diag}(d1_n) - d)_{ij}$.

At last, the Kirchhoff law can be written as $Af + Bd = 0$, where A denotes a linear operator $A : \mathbb{R}^{m \times n} \to \mathbb{R}^{n \times n}$ such that

$$Af = (Nf)^\top, \tag{6}$$

and B denotes a linear operator such that

$$Bd = \operatorname{diag}(d1_n) - d. \tag{7}$$

Bd is also known as a *traffic Laplacian.*

Our final formulation of the *traffic assignment problem* is

$$\Psi(f) = \sum_{e \in \mathcal{E}} \sigma_e \left(\sum_i f_{ei} \right) \to \min_{\substack{Af+Bd=0 \\ f \geq 0}}. \tag{8}$$

According to Theorem 4 from [49], problem (8) has following dual problem

$$Q(t) = \sum_{ij \in OD} d_{ij} T_{ij}(t) - \sum_{e \in \mathcal{E}} \sigma_e^*(t_e) \to \max_{t \geqslant \bar{t}}, \tag{9}$$

where $T_{ij}(t) = T_{p_{ij}}(t) = \min_{p \in P_{ij}} T_p(t)$ is the travel cost from source i to destination j and

$$\sigma_e^*(t_e) = \sup_{f_e \geqslant 0} \{ t_e f_e - \sigma_e(f_e) \} = \bar{f}_e \left(\frac{t_e - \bar{t}_e}{\bar{t}_e \rho} \right)^\mu \frac{(t_e - \bar{t}_e)}{1 + \mu}$$

is the convex conjugate function of $\sigma_e(f_e)$, $e \in \mathcal{E}$.

Subsequently, we can employ the duality gap

$$\Delta(f, t) = \Psi(f) - Q(t) \tag{10}$$

to serve as a convergence measure for traffic assignment problem solvers. It is invariably nonnegative and vanishes only at the solution (f^*, t^*). This quantity is also known as simply *gap* in the traffic assignment literature [10,46,4].

2.3 Trip Distribution. Entropy Model

Trip distribution problem is to find the traffic matrix if total input and output flow is known for each vertex along with travel times between each pair of vertices. Now, we derive the entropy-based single mode trip distribution model of [59]. The basic assumption is that the probability of the distribution d_{ij} is proportional to the number of states of the system with particular distribution, and which satisfy the constraints

$$Q(l,w) = \Big\{d \geq 0 \mid \sum_j d_{ij} = l_i, \sum_i d_{ij} = w_j\Big\}. \tag{11}$$

By defining $h = \begin{pmatrix} l \\ w \end{pmatrix}$ we can rewrite $Q(l,w)$ via an affine constraint as

$$Q(l,w) = \Big\{d \geq 0 \mid Kd = h\Big\}, \tag{12}$$

where linear operator K acts as

$$Kd = \begin{pmatrix} d1_n \\ d^\top 1_n \end{pmatrix}. \tag{13}$$

Suppose

$$D = \sum_i l_i = \sum_j w_j \tag{14}$$

is the total number of trips. The number of distinct arrangements of agents $W(d_{ij})$, which produce distribution d_{ij}

$$W(d_{ij}) = \frac{D!}{d_{11}!(D-d_{11})!} \cdot \frac{(D-d_{11})!}{d_{12}!(D-d_{11}-d_{12})!} \cdots = \frac{D!}{\prod_{ij} d_{ij}!}. \tag{15}$$

To find the most probable distribution of trips d_{ij} we solve

$$\log W(d_{ij}) \to \max_{d \in Q(l,w)}, \tag{16}$$

sing Stirling's approximation (i.e., $\log D! = D\log D - D$), the objective function can be written as

$$\log W(d_{ij}) = \log D! - \sum_{ij} \log d_{ij}! \approx \log D! - \sum_{ij} d_{ij}\log d_{ij} + \sum_{ij} d_{ij}. \tag{17}$$

Taking into account that $\log D!$ and $\sum_{ij} d_{ij} = D$ are constants, the problem can be rewritten as

$$\sum_{ij} d_{ij}\log d_{ij} \to \min_{d \in Q(l,w)} \tag{18}$$

It is also common to assume that larger travel time between two nodes corresponds to lower trip rate between them. Accounting for the travel time factor we obtain

$$\sum_{ij} d_{ij}\log d_{ij} - \gamma \sum_{ij} T_{ij} d_{ij} \to \min_{d \in Q(l,w)}, \tag{19}$$

where $\gamma > 0$ is a scalar parameter, and T_{ij} is the generalized cost of single trip from $i \in \mathcal{V}$ to $j \in \mathcal{V}$, in the simplest case defined as the shortest path travel time between i and j. [59] justifies the addition of the travel time term by interpreting it as the Lagrangian relaxation of the "budget constraint" $\sum_{ij} T_{ij} = T$. Another argument for incorporating the budget constraint is to maintain correspondence with the well-known gravity model: entropy model and gravity model are equivalent when the exponential gravity function is utilized.

Note, that to obtain T_{ij} one needs to solve the traffic assignment problem, and the traffic assignment problem clearly requires the traffic matrix d as its parameter. This is the reason why CDA is usually solved in practice by sequentially solving the two subproblems.

2.4 CDA

CDA is to find such pair (f, d), that d is the solution of entropy model subproblem, where travel times are calculated w.r.t. f, and f is the solution of equilibrium traffic assignment subproblem with traffic matrix being d [8]. This fixed point problem can also be formulated as a convex optimization problem.

Since trip distribution problem (18) and dual traffic assignment problem (9) share the only term $d_{ij}T_{ij}$ which couples travel times t and trip rates d, solution of the problem

$$S(t,d) = \frac{1}{\gamma}\sum_{ij} d_{ij}\log d_{ij} - \sum_{ij} T_{ij}d_{ij} + \sum_{e\in\mathcal{E}} \sigma_e^*(t_e) \to \min_{\substack{t\geq\bar{t} \\ d\in Q(l,w)}},$$

is the solution to both (18) and (9). By the Lagrangian duality we obtain that problem

$$\begin{aligned} P(f,d) = \sum_e \sigma_e(f_e) + \frac{1}{\gamma}\sum_{ij} d_{ij}\ln d_{ij} \to \min_{f,d} \\ \text{s.t. } Af + Bd = 0,\ Kd = h, \\ f \geq 0,\ d \geq 0. \end{aligned} \tag{20}$$

combines both trip distribution problem (18) and (primal) traffic assignment problem (8) [25]. Problem (20) is our final formulation of CDA.

3 Application of Saddle-Point Algorithms

Our formulation (20) of CDA is an affine-constrained convex optimization problem with matrix variables of reasonable dimension (typically, $n \sim 10^3$, $m \sim 10^4$–10^5 in detailed transportation models of large cities). This formulation allows to employ any black-box algorithm for convex optimization with affine constraints. We propose to solve CDA via an optimal first-order algorithm for strongly convex optimization under affine constraints [54], which we will call here OFAC for short. This is an accelerated variant of Proximal Alternating Predictor-Corrector

(PAPC) algorithm which employs Nesterov's and Chebyshev's acceleration mechanisms to achieve convergence rates which match lower bounds for this class of problems. Nesterov's acceleration is used to improve the dependency on the objective's condition number, while Chebyshev's acceleration allows to separate gradient calculation and matrix multiplication complexities and improve the dependency on the affine constraints' condition number. Algorithm 1 and Algorithm 2 provide listings of the OFAC algorithm and its Chebyshev's acceleration subroutine respectively.

Algorithm 1 OFAC

Input data: $x^0 \in X, \mathbf{M}, b \in \operatorname{Im} \mathbf{M}$, Proj_X, ∇F.
Parameters: $\lambda_1 \geq \sigma_{\max}^2(\mathbf{M})$, $\lambda_2 \leq {\sigma_{\min}^{+}}^2(\mathbf{M})$, $L \geq \max_{x \in X} \lambda_{\max}\left(\nabla^2 F(x)\right)$,
    $\mu \leq \min_{x \in X} \lambda_{\min}\left(\nabla^2 F(x)\right)$
$\tau := \sqrt{\frac{19\mu}{44L}}$, $\eta := \frac{1}{4\tau L}$, $\theta := \frac{15}{19\eta}$, $\alpha := \mu$,
$x_f^0 := x^0$, $u^0 := 0_X$
for $k = 0, 1, \ldots$ **do**
    $x_g^k := \tau x^k + (1-\tau) x_f^k$
    $x^{k+\frac{1}{2}} := \operatorname{Proj}_X\left((1+\eta\alpha)^{-1}\left(x^k - \eta(\nabla F(x_g^k) - \alpha x_g^k + u^k)\right)\right)$
    $r^k := \theta\left(x^{k+\frac{1}{2}} - \text{Chebyshev}(x^{k+\frac{1}{2}}, \mathbf{M}, b, \lambda_1, \lambda_2)\right)$
    $u^{k+1} := u^k + r^k$
    $x^{k+1} := \operatorname{Proj}_X\left(x^{k+\frac{1}{2}} - \eta(1+\eta\alpha)^{-1} r^k\right)$
    $x_f^{k+1} := x_g^k + \frac{2\tau}{2-\tau}(x^{k+1} - x^k)$
end for

Algorithm 2 Chebyshev iteration

Input data: $z^0 \in X, \mathbf{M}, b \in \operatorname{Im} \mathbf{M}$
Parameters: $\lambda_1 \geq \sigma_{\max}^2(\mathbf{M})$, $\lambda_2 \leq {\sigma_{\min}^{+}}^2(\mathbf{M})$.
$N := \left\lceil \sqrt{\frac{\lambda_1}{\lambda_2}} \right\rceil$, $\rho := \left(\lambda_1 - \lambda_2\right)^2/16$, $\nu := (\lambda_1 + \lambda_2)/2$
$\gamma^0 := -\nu/2$
$p^0 := -\mathbf{M}^\top(\mathbf{M} z^0 - b)/\nu$
$z^1 := z^0 + p^0$
for $i = 1, \ldots, N-1$ **do**
    $\beta^{i-1} := \rho/\gamma^{i-1}$
    $\gamma^i := -(\nu + \beta^{i-1})$
    $p^i := \left(\mathbf{M}^\top(\mathbf{M} z^i - b) + \beta^{i-1} p^{i-1}\right)/\gamma^i$
    $z^{i+1} := z^i + p^i$
end for
Output: z^N

Clearly, we can apply OFAC to the problem (20) by denoting $x = \begin{pmatrix} f \\ d \end{pmatrix}$, $F(x) = P(f, d)$, $X = \mathbb{R}_+^{(m+n)\times n}$, $\mathbf{M} = \begin{pmatrix} A & B \\ 0_{n\times m} & K \end{pmatrix}$ and $b = \begin{pmatrix} 0_n \\ h \end{pmatrix}$.

For actual implementation of OFAC we need to derive formulas for multiplication by $\mathbf{M}^\top$ and specify the values of parameters $\lambda_1, \lambda_2, L, \mu$. The formulas for other scalar parameters in Algorithm 1 and the formula for N in Algorithm 2 are taken from Theorem 2 [54] and Section 6.3.1 [54] respectively. Note, that the reference has a typo in the formula $\tau = \min\left\{1, \frac{1}{2}\sqrt{\frac{19}{15\kappa}}\right\}$. 19/15 should be 19/11, as stated in Section 6.3.1, and because $\kappa = L/\mu \geq 1$, we get $\tau = \sqrt{\frac{19\mu}{44L}}$.

Since $Af = (Nf)^\top$, and by the definition of adjoint operator $A^\top$ it must holds $\langle y, Af\rangle = \langle y, (Nf)^\top\rangle = \langle yN, f^\top\rangle = \langle (yN)^\top, f\rangle = \langle A^\top y, f\rangle$, we obtain

$$A^\top y = (yN)^\top. \tag{21}$$

Similarly, $\langle y, Bd\rangle = \langle y, \mathrm{diag}(d1_n) - d\rangle = \langle \mathrm{diag}(y)1_n^\top - y, d\rangle = \langle B^\top y, d\rangle$ and

$$B^\top y = \mathrm{diag}(y)1_n^\top - y.$$

Finally, denoting $y = (y_l^\top \; y_w^\top)^\top$ we get $\langle y, Kd\rangle = \langle y_l, d1_n\rangle + \langle y_w, d^\top 1_n\rangle = \langle y_l 1_n^\top, d\rangle + \langle y_w 1_n^\top, d^\top\rangle = \langle y_l 1_n^\top + 1_n y_w^\top, d\rangle = \langle K^\top y, d\rangle$ and

$$K^\top y = y_l 1_n^\top + 1_n y_w^\top.$$

To choose the values of parameters λ_1, λ_2 we need to estimate the maximal and the minimal positive singular values of $\mathbf{M}$.

Lemma 1. *For K defined by* (13) *maximal and minimal positive singular values are given by $\sigma^2_{\max}(K) = 2n$, ${\sigma^+_{\min}}^2(K) = n$.*

Proof.

$$\sigma^2_{\max}(K) := \max_{\|d\|\leq 1} \|Kd\|^2. \tag{22}$$

This is a convex optimization problem, and by Karush-Kuhn-Tucker's theorem [33], [36] d^* is its solution iff (d^*, λ^*) is a stationary point of the Lagrangian

$$\mathcal{L}(d, \lambda) = \|d1_n\|^2 + \|d^\top 1_n\|^2 + \lambda\left(\|d\|^2 - 1\right).$$

For $(d^*, \lambda^*) = (1_n 1_n^\top/n, -2)$ we have

$$\nabla_d \mathcal{L}(d^*, \lambda^*) = 2(d^* 1_n 1_n^\top + 1_n 1_n^\top d^* + \lambda^* d^*) = 2(2d^* + \lambda^* d^*) = 0_{n\times n},$$
$$\nabla_\lambda \mathcal{L}(d^*, \lambda^*) = \|d^*\|^2 - 1 = 0.$$

Thus (d^*, λ^*) is a stationary point of $\mathcal{L}$ and maximum in (22) is attained at d^*:

$$\sigma^2_{\max}(K) = \|Kd^*\|^2 = \|d^* 1_n\|^2 + \|{d^*}^\top 1_n\|^2 = 2n. \tag{23}$$

The minimal positive singular value is defined as

$$ {\sigma^+_{\min}}^2(K) := \min_{d \in \ker^\perp K} \frac{\|Kd\|^2}{\|d\|^2}. \tag{24} $$

It is well known that for a real-valued matrix K the orthogonal complement to its kernel is equal to the image of its transpose: $\ker^\perp K = \operatorname{Im} K^\top$. Thus, by (3)

$$ {\sigma^+_{\min}}^2(K) = \min_y \frac{\|KK^\top y\|^2}{\|K^\top y\|^2} = \min_y \frac{\|K(y_l 1_n^\top + 1_n y_w^\top)\|^2}{\|(y_l 1_n^\top + 1_n y_w^\top)\|^2}. \tag{25} $$

Define $s_w = 1_n^\top y_w,\ s_l = 1_n^\top y_l$, then

$$ \begin{aligned} K(y_l 1_n^\top + 1_n y_w^\top) &= \begin{pmatrix} y_l 1_n^\top 1_n + 1_n y_w^\top 1_n \\ 1_n y_l^\top 1_n + y_w 1_n^\top 1_n \end{pmatrix} = \begin{pmatrix} n y_l + 1_n s_w \\ n y_w + 1_n s_l \end{pmatrix}, \\ \|K(y_l 1_n^\top + 1_n y_w^\top)\|^2 &= n^2 \|y\|^2 + 4 n s_w s_l + n(s_w^2 + s_l^2), \\ \|y_l 1_n^\top + 1_n y_w^\top\|^2 &= \|y_l 1_n^\top\|^2 + 2\langle y_l 1^\top, 1 y_w^\top\rangle + \|1_n y_w^\top\|^2 = n\|y\|^2 + 2 s_l s_w, \end{aligned} $$

Substituting this into the (25) we obtain

$$ {\sigma^+_{\min}}^2(K) = n \frac{n\|y\|^2 + 4 s_l s_w + (s_w^2 + s_l^2)}{n\|y\|^2 + 2 s_l s_w} \geq n, $$

and the inequality becomes equality e.g. when $s_w = s_l = 0$. □

The flattened representation of $Af + Bd$ can be obtained by multiplying flattened representations of f and d by a block-diagonal matrix:

$$ \mathrm{Flatten}(Af + Bd) = \begin{pmatrix} N f_{\bullet 1} + \left[(1_n\ 0_n \dots 0_n)^\top - I_n\right] d_{1\bullet} \\ \vdots \\ N f_{\bullet n} + \left[(0_n\ \dots 0_n 1_n)^\top - I_n\right] d_{n\bullet} \end{pmatrix} $$

where $X_{\bullet i}, X_{i \bullet}$ are column vectors made from i-th column and row of X respectively. Therefore

$$ \begin{aligned} \sigma^+_{\min}(A\ B) &= \min_{i=1,\dots,n} \sigma^+_{\min}\left(N\ \ \left[e_i^\top \otimes 1_n - I_n\right]\right), \\ \sigma_{\max}(A\ B) &= \max_{i=1,\dots,n} \sigma_{\max}\left(N\ \ \left[e_i^\top \otimes 1_n - I_n\right]\right), \end{aligned} \tag{26} $$

what can be computed numerically via SVD.

To finish the derivation of λ_1, λ_2 we will use the fact that for a vertical block matrix $M = \begin{pmatrix} M_1 \\ \vdots \\ M_k \end{pmatrix}$ it holds that $\sigma_{\max}(M) \leq \sqrt{\sum_{i=1}^k \sigma^2_{\max}(M_i)}$. We can also expect $\sigma^+_{\min}(M)$ to be of the same order as $\min_{i=1,\dots,k}\{\sigma^+_{\min}(M_i)\}$. Thus

$\sigma_{\max}(\mathbf{M}) \leq \sqrt{4n^2 + \sigma_{\max}^2(A\ B)}\}$ and $\sigma_{\min}^+(\mathbf{M}) \sim \min\{n, \sigma_{\min}^+(A\ B)\}$, where $\sigma_{\min}^+(A\ B)$ and $\sigma_{\max}(A\ B)$ can be computed via SVD using (26).

Unfortunately, we cannot a priori estimate strong convexity and Lipschitz smoothness constants μ and L since the objective $P(f,d)$ is not Lispschitz smooth nor strongly convex on its domain. Therefore, convergence properties of a first-order method are defined by local convexity and smoothness properties of the objective in a region around the solution. In this situation we resort to some manual tweaking of the parameters L, μ or a grid search.

4 Experiments

To evaluate the performance of OFAC we plot its convergence curve against the most popular algorithm to date [53] — the sequential procedure, which is realized by sequentially solving the trip distribution problem (18) and the traffic assignment problem (8) in a loop. The sequential procedure requires an inertia-like feedback mechanism for convergence [53,47]. For that, we average travel times T_{ij} with their values from the previous iteration before feeding them into the trip distribution subproblem (18).

The traffic assignment subproblem in the sequential procedure is solved by a 3-conjugate Frank-Wolf (3CFW) [32] algorithm, which is a variant of the conjugate Frank-Wolf algorithm [46], that showed the best performance on the given network, comparing with other Frank-Wolf algorithm's modifications. 3CFW is terminated then the duality gap (10) meets the threshold. We plot convergence curves for various values of the threshold, see "Gap" values at the legend of Figure 1. The trip distribution subproblem is solved via Sinkhorn algorithm, which converges to very high precision and its runtime is negligible comparing to that of 3CFW.

The convergence plots are shown in Figure 1. We used a well-known [46,43,47] Sioux-Falls transportation network [57] with $n = 24$ nodes and $m = 76$ edges. To estimate the distance to the solution, OFAC was run with an increased runtime to obtain an approximate solution which is several orders more accurate according to the primal-dual optimality conditions. We choose the distance to the solution as the metric because OFAC and Sinkhorn are primal-dual algorithms that violate the constraints during the optimization process and satisfy them precisely only asymptotically. Consequently, the function value at some iteration can be strictly less than the optimal function value. In such cases, we believe that distance to the solution is the most sensible and comprehensible metric to use.

From Figure 1 we see that OFAC converges much faster and does not suffer from stucking on a limited precision determined by the method parameter values.

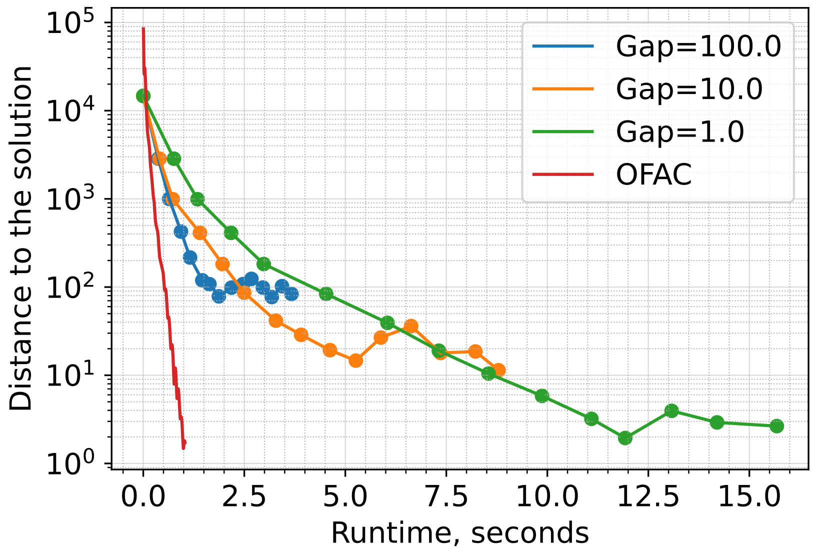

Fig. 1. Comparison of OFAC and the sequential procedure on the Sioux-Falls [57] network

Since we plot the convergence metric against the runtime, it is important to discuss the implementation details. In both approaches, the Python language is used to implement the high-level algorithm logic and data processing, while all computationally expensive calculations are performed by efficient open-source libraries implemented in C/C++. A high-performance implementation (in C++) of Dijkstra algorithm from the graph-tool library [52] is used for shortest path finding in 3CFW, and all other operations with vectors and matrices are done via the NumPy framework [27], whose backend is written in C. For large networks, GPU acceleration of OFAC can be done by a drop-in replacement of NumPy with the jax.numpy module [14]. Unfortunately, on larger networks, the OFAC algorithm, designed for strongly convex functions, exhibits poor convergence due to the high variation of edge flow density in the solution, which means very poor conditioning of the objective function. Further research is required to design a first order primal-dual algorithm with adaptivity and/or preconditioning properties to enhance the efficiency of our approach on large networks. The code is available on GitHub[4].

5 Conclusion

We propose a new algorithmic approach for solving the combined transportation model, offering multiple implementation and computational advantages.

[4] https://github.com/niquepolice/mmo_tm/tree/sp-combined

Promising convergence properties have been demonstrated through numerical experiments on a real network. However, further research into the application of adaptive preconditioning and the exploitation of other problem-specific features is necessary for practical use and scalability to larger networks.

Acknowledgments. The work of Yarmoshik D. is supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) No. 075-03-2024-117, project No. FSMG-2024-0025. The work of Persiianov M. is supported by the annual income of the Endowment Fund of Moscow Institute of Physics and Technology (target capital no. 5 for the development of artificial intelligence and machine learning, https://fund.mipt.ru/capitals/ck5/).

References

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows (1988)
2. Ali Safwat, K.N., Magnanti, T.L.: A combined trip generation, trip distribution, modal split, and trip assignment model. Transportation Science **22**(1), 14–30 (1988)
3. Bar-Gera, H.: Origin-based algorithm for the traffic assignment problem. Transportation Science **36**(4), 398–417 (2002)
4. Bar-Gera, H., Boyce, D.: Origin-based algorithms for combined travel forecasting models. Transportation Research Part B: Methodological **37**(5), 405–422 (2003)
5. Bar-Gera, H., Boyce, D.: Solving a non-convex combined travel forecasting model by the method of successive averages with constant step sizes. Transportation Research Part B: Methodological **40**(5), 351–367 (2006)
6. Beckmann, M., McGuire, C.B., Winsten, C.B.: Studies in the economics of transportation. Tech. rep. (1956)
7. Boyce, D.: Is the sequential travel forecasting paradigm counterproductive? Journal of urban planning and development **128**(4), 169–183 (2002)
8. Boyce, D.: Network equilibrium models for urban transport. Handbook of regional science pp. 247–275 (2021)
9. Boyce, D., Bar-Gera, H.: Multiclass combined models for urban travel forecasting. Networks and Spatial Economics **4**, 115–124 (2004)
10. Boyce, D., Ralevic-Dekic, B., Bar-Gera, H.: Convergence of traffic assignments: how much is enough? Journal of Transportation Engineering **130**(1), 49–55 (2004)
11. Boyce, D.E.: Urban transportation network-equilibrium and design models: recent achievements and future prospects. Environment and planning A **16**(11), 1445–1474 (1984)
12. Boyce, D.E., Zhang, Y.F., Lupa, M.R.: Introducing" feedback" into four-step travel forecasting procedure versus equilibrium solution of combined model. Transportation Research Record **1443**, 65 (1994)
13. Boyles, S.D., Lownes, N.E., Unnikrishnan, A.: Transportation Network Analysis, vol. 1 (2023), edition 0.91
14. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018), http://github.com/google/jax

15. Chen, A., Yang, C., Kongsomsaksakul, S., Lee, M.: Network-based accessibility measures for vulnerability analysis of degradable transportation networks. Networks and Spatial Economics **7**, 241–256 (2007)
16. Cheng, L., Du, M., Jiang, X., Rakha, H.: Modeling and estimating the capacity of urban transportation network with rapid transit. Transport **29**(2), 165–174 (2014)
17. de Dios Ortúzar, J., Willumsen, L.G.: Modelling transport. John wiley & sons (2011)
18. Du, M., Jiang, X., Cheng, L.: Alternative network robustness measure using system-wide transportation capacity for identifying critical links in road networks. Advances in Mechanical Engineering **9**(4), 1687814017696652 (2017)
19. Evans, S.P.: Derivation and analysis of some models for combining trip distribution and assignment. Transportation research **10**(1), 37–57 (1976)
20. Florian, M., Nguyen, S.: A combined trip distribution modal split and trip assignment model. Transportation Research **12**(4), 241–246 (1978)
21. Florian, M., Nguyen, S., Ferland, J.: On the combined distribution-assignment of traffic. Transportation Science **9**(1), 43–53 (1975)
22. Frank, M., Wolfe, P., et al.: An algorithm for quadratic programming. Naval research logistics quarterly **3**(1-2), 95–110 (1956)
23. Friesz, T.L.: An equivalent optimization problem for combined multiclass distribution, assignment and modal split which obviates symmetry restrictions. Transportation Research Part B: Methodological **15**(5), 361–369 (1981)
24. Garrett, M., Wachs, M.: Transportation planning on trial: The Clean Air Act and travel forecasting. Sage Publications (1996)
25. Gasnikova, E., Gasnikov, A., Yarmoshik, D., Kubentaeva, M., Persiianov, M., Podlipnova, I., Kotlyarova, E., Sklonin, I., Podobnaya, E., Matyukhin, V.: Sufficient conditions for multi-stages traffic assignment model to be the convex optimization problem. arXiv preprint arXiv:2305.09069 (2023)
26. Guminov, S., Dvurechensky, P., Tupitsa, N., Gasnikov, A.: On a combination of alternating minimization and nesterov's momentum. In: International Conference on Machine Learning. pp. 3886–3898. PMLR (2021)
27. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2, https://doi.org/10.1038/s41586-020-2649-2
28. Ho, H., Wong, S.: Housing allocation problem in a continuum transportation system. Transportmetrica **3**(1), 21–39 (2007)
29. Ho, H., Wong, S., Loo, B.P.: Combined distribution and assignment model for a continuum traffic equilibrium problem with multiple user classes. Transportation Research Part B: Methodological **40**(8), 633–650 (2006)
30. Horowitz, A.J.: Tests of an ad hoc algorithm of elastic-demand equilibrium traffic assignment. Transportation Research Part B: Methodological **23**(4), 309–313 (1989)
31. Huang, H.J., Lam, W.H.: Modified evans' algorithms for solving the combined trip distribution and assignment problem. Transportation Research Part B: Methodological **26**(4), 325–337 (1992)
32. Ignashin, I., Yaramoshik, D.: Modifications of the frank-wolfe algorithm in the problem of finding the equilibrium distribution of traffic flows. Mathematical Modeling and Numerical Simulation **10**(1), 10–25 (2024)

33. Karush, W.: Minima of functions of several variables with inequalities as side constraints. M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago (1939)
34. Kim, Y., Samaranayake, S., Wischik, D.: A combined convex model for travel demand forecasting with hierarchical extended logit model. arXiv preprint arXiv:2308.01817 (2023)
35. Kubentayeva, M., Yarmoshik, D., Persiianov, M., Kroshnin, A., Kotliarova, E., Tupitsa, N., Pasechnyuk, D., Gasnikov, A., Shvetsov, V., Baryshev, L., et al.: Primal-dual gradient methods for searching network equilibria in combined models with nested choice structure and capacity constraints. Computational Management Science **21**(1), 15 (2024)
36. Kuhn, H.W., Tucker, A.W.: Nonlinear programming, in (j. neyman, ed.) proceedings of the second berkeley symposium on mathematical statistics and probability (1951)
37. Lam, W.H., Huang, H.J.: A combined trip distribution and assignment model for multiple user classes. Transportation Research Part B: Methodological **26**(4), 275–287 (1992)
38. Larsson, T., Patriksson, M.: Simplicial decomposition with disaggregated representation for the traffic assignment problem. Transportation Science **26**(1), 4–17 (1992)
39. LeBlanc, L.J., Farhangian, K.: Efficient algorithms for solving elastic demand traffic assignment problems and mode split-assignment problems. Transportation Science **15**(4), 306–317 (1981)
40. Lee, D.H., Wu, L., Meng, Q.: Equity based land-use and transportation problem. Journal of Advanced Transportation **40**(1), 75–93 (2006)
41. Leventhal, T., Nemhauser, G., Trotter Jr, L.: A column generation algorithm for optimal traffic assignment. Transportation Science **7**(2), 168–176 (1973)
42. Lin, J.J., Feng, C.M.: A bi-level programming model for the land use–network design problem. The annals of regional science **37**, 93–105 (2003)
43. Liu, Z., Yin, Y., Bai, F., Grimm, D.K.: End-to-end learning of user equilibrium with implicit neural networks. Transportation Research Part C: Emerging Technologies **150**, 104085 (2023)
44. Lundgren, J.T., Patriksson, M.: An algorithm for the combined distribution and assignment model. In: Transportation Networks: Recent Methodological Advances. Selected Proceedings of the 4th EURO Transportation MeetingAssociation of European Operational Research Societies (1999)
45. McNally, M.G.: The activity-based approach (2000)
46. Mitradjieva, M., Lindberg, P.O.: The stiff is moving—conjugate direction frank-wolfe methods with applications to traffic assignment. Transportation Science **47**(2), 280–293 (2013)
47. Najmi, A.: Interaction of demand and supply in transport planning model systems: A comprehensive revisit. Ph.D. thesis, UNSW Sydney (2020)
48. Nesterov, Y.: Universal gradient methods for convex optimization problems. Mathematical Programming **152**(1), 381–404 (2015). https://doi.org/10.1007/s10107-014-0790-0, http://dx.doi.org/10.1007/s10107-014-0790-0
49. Nesterov, Y., De Palma, A.: Stationary dynamic solutions in congested transportation networks: summary and perspectives. Networks and spatial economics **3**, 371–395 (2003)
50. Oppenheim, N.: Equilibrium trip distribution/assignment with variable destination costs. Transportation Research Part B: Methodological **27**(3), 207–217 (1993)
51. Oppenheim, N., et al.: Urban travel demand modeling: from individual choices to general equilibrium. John Wiley and Sons (1995)

52. Peixoto, T.P.: The graph-tool python library. figshare (2014). https://doi.org/10.6084/m9.figshare.1164194, http://figshare.com/articles/graph_tool/1164194
53. Reeder, P., Bhat, C., Lorenzini, K., Hall, K., et al.: Positive feedback: exploring current approaches in iterative travel demand model implementation. (2012)
54. Salim, A., Condat, L., Kovalev, D., Richtárik, P.: An optimal algorithm for strongly convex minimization under affine constraints. In: International Conference on Artificial Intelligence and Statistics. pp. 4482–4498. PMLR (2022)
55. Sheffi, Y.: Urban transportation networks, vol. 6. Prentice-Hall, Englewood Cliffs, NJ (1985)
56. Tam, M., Lam, W.H.: Maximum car ownership under constraints of road capacity and parking space. Transportation Research Part A: Policy and Practice **34**(3), 145–170 (2000)
57. Transportation Networks for Research Core Team: Transportation networks for research. https://github.com/bstabler/TransportationNetworks (2024), accessed: 2024-02-29
58. US Bureau of Public Roads: Traffic Assignment Manual. Department of Commerce, Urban Planning Division, Washington D.C. (1964)
59. Wilson, A.G.: The use of entropy maximising models, in the theory of trip distribution, mode split and route split. Journal of transport economics and policy pp. 108–126 (1969)
60. Wong, K.I., Wong, S., Wu, J., Yang, H., Lam, W.H.: A combined distribution, hierarchical mode choice, and assignment network model with multiple user and mode classes. Urban and regional transportation modeling pp. 25–42 (2004)
61. Xu, M., Chen, A., Gao, Z.: An improved origin-based algorithm for solving the combined distribution and assignment problem. European Journal of Operational Research **188**(2), 354–369 (2008)
62. Yang, H., Bell, M.G., Meng, Q.: Modeling the capacity and level of service of urban transportation networks. Transportation Research Part B: Methodological **34**(4), 255–275 (2000)
63. Yim, K.K., Wong, S., Chen, A., Wong, C.K., Lam, W.H.: A reliability-based land use and transportation optimization model. Transportation Research Part C: Emerging Technologies **19**(2), 351–362 (2011)